Volume 14, No. 1
January 2010

Milam Aiken is a Professor and Chair of Management Information Systems in the School of Business Administration at the University of Mississippi. He has been ranked as a leading researcher in the field of Group Support Systems.
Milam Aiken can be reached at maiken@bus.olemiss.edu.

Mina Park is an assistant professor of Management Information Systems in the School of Business at Northern State University, South Dakota. Her primary research interests are Group Support Systems, and she is currently working on various multilingual electronic meeting studies.
Mina Park can be reached at Mina.Park@northern.edu.

Front Page

Select one of the previous 50 issues.
Select an issue:

The Efficacy of Round-trip Translation for MT Evaluation

by Milam Aiken
School of Business Administration
University of Mississippi

Mina Park
School of Business
Northern State University

Introduction

Round-trip translation (RTT), otherwise known as reverse translation or back-and-forth translation, involves the translation of text from one language to another (the forward translation or FT) and back again (the back translation or BT), e.g., a paragraph written in English can be translated to Spanish, and then, the resulting Spanish text can be translated back to English again. RTT has been used in several previous studies (e.g., Glidden-Tracey & Greenwood, 1997; Klaudy, 1996), as some have argued that if the final and original texts are identical or very similar, RTT provides strong evidence that the forward translation is of high quality (Chan, 2006).

However, others have severely rebuked RTT as a valid technique for machine translation (MT) evaluation. For example, van Zaanen and Zwarts (2006) state:

> "Lay people discussing machine translation systems often perform a round trip translation, that is translating a text into a foreign language and back, to measure the quality of the system. The idea behind this is that a good system will produce a round trip translation that is exactly (or perhaps very close to) the original text. However, people working with machine translation systems intuitively know that round trip translation is not a good evaluation method. In this article we will show empirically that round trip translation cannot be used as a measure of the quality of a machine translation system. Even when using translations of multiple machine translation systems into account, to reduce the impact of errors of a single system, round trip translation cannot be used to measure machine translation quality."

O'Connell (2001) asserts:

> "A common misunderstanding about MT evaluation is the belief that back translation can disclose a system's usability. ... The theory is that if back translation returns [the source language] input exactly, the system performs well for this language pair."

Somers (2006) adds:

> "RTT (grunt), what is it good for? Absolutely nothing."

Crystal (2004) writes:

> "This makes it hard to believe that any back translation would be a reliable method of verifying the accuracy of a translation ... We believe that back translation provides absolutely nothing

> syntactically or semantically, about the translation and is unreliable as an effective quality control procedure in translation."

However, Crystal (2004) then contradicts the statement above by asserting that RTT might sometimes be acceptable:

> "Technical documents, like MSDS or scientific formulas seem to be suitable for back translation because the source text is usually written by Engineers or Scientist and is less likely to include humour, colloquial expressions or complex literary statements. If the content seems like it is written by a computer, then it is easier to obtain an verbatim translation in the target language and a back translation would be helpful to verify the content. ... RTT might be used if there is no other viable method to determine the accuracy of the work of a translation, e.g. no equivalent text or no human fluent in the target language is available."

Some (e.g., Yamashita & Ishida, 2006) have even gone so far as to claim that translations between two languages are not transitive (i.e., language A to language B then B to A does not yield the original expression in language A). However, this is not always the case, as the following RTT examples using the free online translation service *Google Translate* (http://translate.google.com/) prove:

English: How are you?

German: Wie geht es dir?

English: How are you?

English: What is your name?

Spanish: ¿Cómo te llamas?

English: What is your name?

English: Who are you?

Danish: Hvem er du?

English: Who are you?

Sometimes, the final result of RTT is not exact, but has the same meaning:

English: Where is the store?

Dutch: Waar is de winkel?

English: Where is the shop?

As these examples show, at least sometimes, RTT can give accurate results (i.e., a good forward translation is associated with a good back translation). The problem with using RTT for MT evaluations is that a bad round trip might be due to a poor back translation of what was a perfectly reasonable forward translation, or vice versa, or both (Somers, 2007). Furthermore, a good round trip might mask a poor outward translation Evaluators cannot tell if errors occurred during the passage to the target language or during the return passage to the source language, and any errors that occur in the first translation could cause more problems in the back translation.

Previous research

In one study that tested the efficacy of RTT (Somers, 2006), four sets of text about 100 sentences each representing various language pairs were used with five free online translation services:

1. Babelfish: http://babelfish.yahoo.com
2. Freetranslation: www.freetranslation.com
3. Systran: http://www.systranet.com/

4. ProMT: www.online-translator.com
5. Worldlingo: http://www2.worldlingo.com

The study concluded that RTT is not a particularly good way to identify which system is better. For example, the system with the highest FT score ranked 4th or 5th using the BT score. The BT score was often better than the FT score, and the BT score did not predict a good score for the FT. However, the study used automatic rather human evaluation of translated text.

Even though some studies have reported that BLEU (Papineni, et al., 2002) and F scores (Turian, et al., 2003) are reliable measures of translation accuracy and correlate well with human judgements of quality (e.g., Coughlin, 2003), others have disagreed. For example, during the 2005 NIST MT evaluation, BLEU failed to correspond to the scores produced in the human evaluations (Callison-Burch, et al., 2006). In addition, Huang (1990) writes:

> "Although some have rejected RTT as 'useless', the objections are often based upon use of automatic testing [such as BLEU], idioms, etc. ... So our conclusion is really that RTT cannot tell good MT systems from bad ones, or easy-to-translate texts from hard ones, *based on automatic evaluation methods*."

In addition to the problems that might have occurred using BLEU and F scores, poor results might have been obtained due to the use of rule-based translation systems rather than a statistical-learning approach as *Google Translate* uses.

Finally, at least one study (Shigenobu, 2007) found a correlation between FT and BT accuracies, indicating RTT can be useful.

An RTT analysis

One study (Somers, 2006) concluded that RTT was not reliable, but another (Shigenobu, 2007) found that it was. To investigate the reliability further, we conducted a study with ten text passages from Chall & Dale (1995) that ranged in complexity from 18.4 (hard) to 100 (easy) on the Flesch Reading Ease scale:

1. One morning Toad sat in bed. "I have many things to do," he said. "I will write them all down on a list so that I can remember them." Toad wrote on a piece of paper: A list of things to do today. Then he wrote: Wake up. "I have done that," said Toad, and crossed it out.
2. "You said you didn't want it," said Thelma. "And anyhow, I don't want to sell it now." "Why not?" said Frances. "Well," said Thelma, "it is a very good tea set. It is plastic that does not break. It has pretty red flowers on it. It has all the cups and saucers. It has the sugar bowl and the cream pitcher and the teapot. It is almost new, and I think it cost a lot of money." "I have two dollars and seventeen cents," said Frances. "That's a lot of money." "I don't know," said Thelma. "If I sell you ...
3. Once upon a time a very small witch was walking in the woods. The cold wind was blowing the dry leaves all around her. The little witch was frantically searching for a house for the winter. She could not find one. Suddenly a piece of orange paper, blown by the wind, landed at her feet. She picked it up. The little witch looked closely at the paper and then she said, "I shall make myself a little house from this piece of orange paper." She folded the paper in half. Then she took her scissors (she always carried a pair ...
4. Seals are wonderful divers. Some seals can dive several hundred feet below the surface. On deep dives, they can stay underwater up to 40 minutes without surfacing to breathe. They have special features to help save oxygen on such dives. When seals dive, they stop breathing. For very deep dives, their blood flows to everything except critical organs. Seals can also slow their heart rates, sometimes to one-tenth the rates at the surface. You may wonder how seals avoid the bends on deep dives. The bends are a painful condition. They are caused when nitrogen dissolves in ...
5. Eskimos of Alaska's Arctic north coast have hunted whales for centuries. Survival has depended on killing the 60-foot-long bowhead whales that swim from the Bering Sea to the ice-clogged Beaufort Sea each Spring. The Eskimos' entire way of life has been centered around the hunt. But now that way of life is being threatened by America's need for oil, say

many Eskimos who hunt the whales. Huge amounts of oil may be beneath the Beaufort Sea. And oil companies want to begin drilling this spring. However, many Eskimos say severe storms and ice conditions make drilling dangerous ...

6. Why is it that as soon as "Jingle Bells" starts playing on the radio, otherwise-sane people are driven to extremes to create the Perfect Christmas? Take the case of Maureen McFadden, a Woman's Day editor, who decided to decorate her tree with homemade gingerbread ornaments. "I started late in the evening," she recalled. "And then I knocked the molasses jar on the floor." It was downhill from there. Her cat - long-haired, of course - sat in the molasses pool. "And when I yelped, he ran down the hall into my bedroom spewing molasses everywhere." Still, after she washed the ...
7. The controversy over the laser-armed satellite boils down to two related questions: Will it be technically effective? And should the United States make a massive effort to deploy it? To its backers, the laser seems the perfect weapon. Traveling in a straight line at 186,000 miles per second, a laser beam is tens of thousands of times as fast as any bullet or rocket. It could strike its target with a power of many watts per square inch. The resulting heat, combined with a mechanical shock wave created by recoil as surface layers were blasted away, would quickly melt ...
8. The latest finding is a refinement of evidence presented last summer by audio expert James Barger - who testified there was a 50 percent probability that four shots were recorded on the tape. Barger had recorded test firings at various points in the Dealey Plaza, then compared them with the motorcycle recording. The greatest similarity was produced by two shots from the book depository, one from the knoll and another from the depository. But Barger did not draw firm conclusions because he could not pin-point the policeman's motorcycle; his estimate could have been 18 feet off in any direction. Weiss, whose ...
9. Until the 1940's, there were no specific psychiatric drugs. Bromides, barbiturates, and opiates were known to sedate disturbed patients but did not reverse the symptoms of severe mental illnesses such as the schizophrenias or manic-depressive psychoses. They did ameliorate anxiety, but only at the cost of fogging the minds of the recipients, who had to decide between being unhappy and being intoxicated. In the 1950's, the first specific drug appeared: chlorpromazine (trade name Thorazine). It was synthesized when an antihistamine chemical relative was found to sedate surgical patients. However, clinical observations showed that this drug did much more than simply ...
10. Further support for the view that educational expansion would reduce inequalities was derived from the dualistic nature of developing societies. The economic structures of developing societies were said to consist of two sectors: a traditional sector that uses little capital, is relatively unproductive, does not require an educated labor force, and places a great emphasis on subsistence farming, small workshops and small commercial enterprises; and a modern sector that uses advanced technology and capital, is far more productive, and requires a labor force with at least some schooling. Expanding the educational system would qualify more workers for jobs where demands ...

A fluent Korean speaker translated the English text to Korean, and *Google Translate* was used to generate another Korean translation. Then, *Google Translate* was used to translate both sets of Korean text back to English. A native English speaker evaluated the final English text for comprehension on a 0 to 100 scale (100 = best). As an extra test, BLEU scores were calculated for each of the 10 text samples, and a "perfect" BLEU ne score (i.e., the original English text was compared to an exact copy of the text) was also added, with the results below:

	One-way		**RTT**		**Perfect**
	Human	BLEU	Human	BLEU	BLEU

1	90	0.31	95	0.47	0.70
2	80	0.34	90	0.34	0.56
3	75	0.32	45	0.42	0.59
4	75	0.31	80	0.30	0.61
5	95	0.27	90	0.38	0.65
6	40	0.33	60	0.40	0.75
7	60	0.43	80	0.44	0.78
8	55	0.37	55	0.47	0.77
9	100	0.27	95	0.43	0.74
10	85	0.31	70	0.43	0.72
average	76.0	0.33	75.5	0.41	0.69

There was no significant difference in the mean ratings between the Korean-to-English one-way translation (mean = 76.0, Std. Dev. = 17.7) and the English-Korean-English round-trip translation (mean = 75.5, std. dev. = 18.9). Further, there was a significant positive correlation ($R = 0.65$, $p = 0.04$) between the two items, indicating that RTT was a good predictor of one-way accuracy.

However, there was no significant correlation between the BLEU score and the human evaluator's rating for either the RTT text ($R = -0.07$, $p < 0.85$) or the one-way translation samples ($R = -0.31$, $p < 0.39$). There was a significant correlation between the perfect BLEU score and the RTT BLEU score ($R = 0.70$, $p < 0.03$), but not between the perfect and one-way BLEU scores ($R = 0.36$, $p < 0.31$). These poor BLEU correlations might be a result of the relatively few sentences in each text passage (Snover, et al., 2006).

Another RTT analysis

In one study (Aiken & Ghosh, 2009), each of the following sentences or phrases were translated into 32 non-English languages with *Google Translate*.

1. How can we improve the parking problem on campus?
2. sell spots on ebay
3. Before accepting so many students they should make a huge parking garage.
4. I think we should all have to ride bicycles to school and tear up the parking lots and plant trees so that the squirrels will have their habitat back.

5. I think the instructor should drive around and take people to class in his little pickup truck
6. I think the residential college will help eliminate the parking problem for commuters.
7. A parking lot can be created outside campus
8. We need to build a parking garage.
9. Parking garage
10. Students who live within a reasonable distance should ride bikes
11. Demolish condemned buildings on campus and turn them into parking lots
12. get more bikes and unicycles
13. Do not allow freshman to bring cars.
14. I do not think that there is a parking problem.
15. Interconnected tunnels would be fun and cost effective

Then, the text was translated back into English with *Google* and 240 undergraduate business students evaluated the text on a scale of 1 = do not understand at all, 2=mostly do not understand, 3= misunderstand more than understand, 4=no opinion, 5=understand more than misunderstand, 6= understand most of it, and 7=understand very clearly. Roughly eight students evaluated text from each language to avoid duplication.

In another study (Aiken, et al., 2009a), two random sentences were obtained for each of the 32 languages used in the Aiken & Ghosh study from the 15 *Omniglot (http://www.omniglot.com/)* sentences below and translated to English with *Google Translate.* However, no two languages shared the same sentences, and thus, direct comparisons between the languages were more difficult.

1. How much is this?
2. Where is the toilet?
3. Would you write it down?
4. Would you like to dance?
5. Please speak more slowly.
6. Pleased to meet you.
7. My hovercraft is full of eels.
8. One language is never enough.
9. I don't understand.
10. I love you.
11. Please say that again.
12. This gentleman will pay for everything.
13. Where are you from?
14. What's your name?
15. Leave me alone.

In a third ranking study (Aiken, et al., 2009b), the equivalents for the five sentences below for each of the 32 languages were obtained from *Omniglot* and translated to English with *Google Translate.*

1. Pleased to meet you.
2. My hovercraft is full of eels.
3. One language is never enough.
4. I don't understand.
5. I love you.

In each of the latter studies, two objective English speakers evaluated the English translations of the foreign text using the scale:

1. The text is clear, easy to understand and grammatically correct and does not require any corrections.
2. The text contains minor errors such as incorrect prepositions or articles, but is otherwise impeccable.
3. The text is a mixture of minor errors and incorrect terms, but the meaning is still understandable.
4. The text is a mixture of minor errors and incorrect terms, and it takes a definite effort to understand the meaning.
5. The text is incomprehensible gibberish.

Results of the three studies are indicated below:

Aiken & Ghosh, 2009		Aiken, et al., 2009a	Aiken, et al., 2009b

higher score better			lower score better			lower score better	
Italian	5.78		Dutch	1.2		Dutch	1.3
Serbian	5.48		Danish	1.3		Czech	1.4
Russian	5.47		Swedish	1.3		Chinese	1.5
Finnish	5.14		German	1.4		Italian	1.5
Dutch	5.09		Norwegian	1.4		Korean	1.5
Bulgarian	5.03		Slovenian	1.5		Portuguese	1.5
Danish	4.98		Portuguese	1.5		French	1.7
Lithuanian	4.95		Polish	1.6		German	1.7
Ukrainian	4.94		Czech	1.7		Russian	1.7
French	4.91		Croatian	1.8		Slovak	1.7
Latvian	4.9		Bulgarian	1.8		Slovenian	1.7
Swedish	4.9		Slovak	1.9		Danish	1.8
Portuguese	4.89		Russian	1.9		Norwegian	1.8
Norwegian	4.82		French	1.9		Bulgarian	1.9
Catalan	4.77		Romanian	1.9		Finnish	1.9
Chinese	4.65		Filipino	2		Polish	1.9
Polish	4.63		Hebrew	2		Filipino	2

Czech	4.62		Latvian	2		Hebrew	2
Vietnamese	4.6		Korean	2.1		Swedish	2
Korean	4.58		Italian	2.1		Croatian	2.2
German	4.55		Catalan	2.2		Catalan	2.3
Croatian	4.49		Serbian	2.2		Japanese	2.3
Slovak	4.47		Ukrainian	2.3		Serbian	2.4
Hebrew	4.45		Finnish	2.3		Ukrainian	2.4
Greek	4.38		Greek	2.3		Vietnamese	2.4
Slovenian	4.33		Chinese	2.4		Greek	2.5
Indonesian	4.3		Indonesian	2.4		Indonesian	2.5
Romanian	4.27		Hindi	2.5		Romanian	2.6
Arabic	4.15		Vietnamese	2.6		Latvian	3.1
Filipino	3.98		Japanese	2.7		Hindi	3.2
Japanese	3.77		Lithuanian	2.9		Arabic	3.4
Hindi	3.75		Arabic	3.3		Lithuanian	3.4

A Pearson correlation analysis showed a significant negative correlation between the two one-way translation rankings of *Omniglot* text ($R = 0.553$, $p < 0.001$), but these two ranking correlations were not significant at $\alpha = 0.05$ with the RTT ranking (2009a: $R= -0.288$, $p < 0.110$; 2009b: $R = -0.333$, $p < 0.063$). The correlations were negative because in the first study, higher scores were better while in the latter two, lower scores were better. Although they were not highly significant, the language rankings obtained with RTT seemed to correlate well with language rankings from a one-way translation. Further studies of FT and RTT are necessary using the same sets of text with a more complete analysis. In addition, it seems clear that RTT provides some benefit in that it indicates which languages are generally translated well (e.g., Dutch, Danish, and Italian) while others are not (e.g., Arabic, Hindi, and Indonesian).

Conclusion

Somers (2006) states:

> "Although it is widely agreed in the MT community that RTT is a bad technique, and equally widely suggested in the lay community that it is an effective way to evaluate systems, there has been little or no work to demonstrate empirically whether RTT is in fact as misleading as it is claimed."

His study indicated that RTT was not reliable, as the technique gave high BLEU and F-scores when the forward translation was poor. On the other hand, our studies and at least one other (Shigenobu, 2007) indicate that the method has predictive power. More study is required with more human evaluators and additional automated techniques.

RTT is not perfect, but no other evaluation technique is, either. For a single given sentence, we cannot know for sure if a good (or bad) RTT indicates that the FT was good (or bad) or vice versa. But, over the length of a longer text or multiple language pairs, RTT quality might reflect the general quality of the system used. In addition, RTT is the only technique that can be used when no human fluent in the target language or equivalent text is readily available.

References

1. Aiken, M. and Ghosh, K. (2009). Automatic translation in multilingual business meetings. *Industrial Management & Data Systems*, 109(7), 916-925.
2. Aiken, M., Park, M., and Lindblom, T. (2009a). Integrating machine translation with group support systems. Working paper, University of Mississippi.
3. Aiken, M., Park, M., Simmons, L., and Lindblom, T. (2009b). Automatic translation in multilingual electronic meetings. *Translation Journal*, 13(9), July.
4. Callison-Burch, C., Osborne, M. and Koehn, P. (2006) "Re-evaluating the Role of BLEU in Machine Translation Research" in *11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006* pp. 249-256.
5. Chall, J. and Dale, E. (1995). Readability Revisited: The New Dale-Chall Readability Formula, Brookline Books: Cambridge, MA.
6. Chan, S. (2006). A Dictionary of Translation Technology. Chinese University Press. Hong Kong.
7. Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. *MT Summit IX, New Orleans, USA*, 23-27.
8. Crystal, S. (2004). Back translation: Same questions - different continent. *Communicate* (London: Association of Translation Companies) (Winter 2004): p. 5.
9. Glidden-Tracey, C. and Greenwood, A. (1997). A validation study of the Spanish self-directed search using back-translation procedures. *Journal of Career Assessment*, 5(1), 105-113.
10. Huang, X. (1990). A machine translation system for the target language inexpert. *13th International Conference on Computational Linguistics*, COLING-90, Vol. 3, Helsinki, 364-367.
11. Klaudy, K. (1996). Back-translation as a tool for detecting explicitation strategies in translation. In: Klaudy, K., & Lambert, J. & Sohár, A. (eds.) *Translation Studies in Hungary*. Budapest: Scholastica. 99-114.
12. O'Connell, T. (2001). Preparing your web site for machine translation: How to avoid losing (or gaining) something in the translation.
13. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311-318. Evaluation and usability of back translation for intercultural communication. *Human-Computer Interaction* (11), 259-265
14. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, pages 223-231, Cambridge, MA, August.

15. Somers, H. (2006). Round-trip translation: What is It Good for? *Proceedings of the Australasian Language Technology Workshop 2005*.
16. Somers, H. (2007). The use of machine translation by law librarians - a reply to Yates. *Law Library Journal*, 99, 611-619.
17. Turian, J., Shen, L., and Melamed, I. (2003). Evaluation of machine translation and its evaluation. *Proceedings of MT Summit IX*, New Orleans, U.S.A.
18. van Zaanen, M. and Zwarts, S. (2006). Unsupervised measurement of translation quality using multi-engine, bi-directional translation. *AI 2006: Advances in Artificial Intelligence*. Springer: Berlin, 1208-1214.
19. Yamashita, N. and Ishida, T. (2006). Effects of machine translation on collaborative work. *Proceedings of the 2006 20th-anniversary Conference on Computer Supported Cooperative Work*. Banff, Alberta, Canada, 515-524.